ELSEVIER

FEMS Microbiology Letters 174 (1999) 247–250

FEMS MICROBIOLOGY LETTERS

BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences

Tatiana A. Tatusova *, Thomas L. Madden

National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received 18 February 1999; received in revised form 17 March 1999; accepted 18 March 1999

Abstract

'BLAST 2 Sequences', a new BLAST-based tool for aligning two protein or nucleotide sequences, is described. While the standard BLAST program is widely used to search for homologous sequences in nucleotide and protein databases, one often needs to compare only two sequences that are already known to be homologous, coming from related species or, e.g. different isolates of the same virus. In such cases searching the entire database would be unnecessarily time-consuming. 'BLAST 2 Sequences' utilizes the BLAST algorithm for pairwise DNA-DNA or protein-protein sequence comparison. A World Wide Web version of the program can be used interactively at the NCBI WWW site (http://www.ncbi.nlm.nih.gov/gorf/bl2.html). The resulting alignments are presented in both graphical and text form. The variants of the program for PC (Windows), Mac and several UNIX-based platforms can be downloaded from the NCBI FTP site (ftp://ncbi.nlm.nih.gov).

Keywords: Algorithm; Sequence alignment; Software

1. Introduction

BLAST is a rapid sequence comparison tool that uses a heuristic approach to construct alignments by optimizing a measure of local similarity [1,2]. Since BLAST compares protein and nucleotide sequences much faster than dynamic programming methods such as Smith-Waterman and Needleman-Wunsch [3,4], it is widely used for database searches. A number of important scientific contexts, however, involve the comparison of only two sequences and do not require a time-consuming database search. This happens for example, when one compares a series of viral isolates that may differ in only several nucleotide residues per genome. Ongoing projects of sequencing closely related microbial genomes, such as the genomes of *Helicobacter pylori* strains 26695 and J99 [5,6], make this a very common task. To meet these needs we have developed a program that uses the BLAST algorithm to produce reliable sequence alignments, without computer-intensive and time-consuming database searches. The BLAST 2 Sequences program finds multiple local alignments between two sequences, allowing the user to detect homologous protein domains or internal sequence duplications. BLAST 2 Sequences has been very useful for the comparison of homologous genes from complete microbial genomes. Using BLAST 2 Sequences for

* Corresponding author. E-mail: tatiana@ncbi.nlm.nih.gov

0378-1097/99/$20.00
PII: S0378-1097(99)00149-4

nucleotide sequence comparison of different strains or isolates of the same virus offers a convenient strategy to study the genome variations and evolutionary events, such as substitutions, insertions and deletions.

2. Methods

2.1. Algorithm

'BLAST 2 SEQUENCES' is an interactive tool that utilizes the BLAST engine for pairwise DNA-DNA or protein-protein sequence comparison and is based on the same algorithm and statistics of local alignments that have been described earlier [1,2]. The BLAST 2.0 algorithm generates a gapped alignment by using dynamic programming to extend the central pair of aligned residues. The heuristic methods confine the alignments to a predefined region of the path graph. A performance evaluation of the new gapped BLAST algorithm and its comparison to that of the original ungapped BLAST [1] and the Smith-Waterman algorithm [3] have been presented [4].

2.2. User-defined parameters

The 'BLAST 2 SEQUENCES' interface allows the user to perform a series of searches with various parameters. The program can align hundreds of sequences within a reasonable time. Different scoring matrices are provided for protein-protein comparisons; each matrix is most sensitive at finding similarities at a specific evolutionary distance. The default matrix, BLOSUM62 [7] is generally considered to be the best for a wide variety of distances.

Changing the gap existence and extension penalties may change the number and length of gaps in an alignment. There is no analytical formula that determines the 'best' gap values to use, so that one may wish to experiment with values in order to explore more of the alignment 'space'. BLAST 2.0 uses affine gap costs, which assess a score (a+bk) for a gap of length 'k' [8]; long gaps do not cost much more than short ones. Note that only limited values for the gap existence and extension penalties are supported, so that Karlin-Altschul statistics [9] can be applied. The default values of parameters are set up by Javascript function. Selection of the scoring matrix changes the default values of the gap penalties. An incorrect parameter setting returns an error response and brings back the main page, allowing the user to change the selection.

Both protein and nucleotide sequences may contain regions of low compositional complexity, which give spuriously high BLAST scores that reflect compositional bias rather than significant position by position alignment [10]. The SEG program [10] can be used for proteins and the DUST program (Tatusov and Lipman, in preparation) for nucleotides if the 'Filter' parameter is set. One may also wish to view alignments without a complexity filter, especially if it seems possible that an important part of the aligned sequence has been filtered over. In that case the dot-plot display (see Fig. 1) notes the locations that would have been masked.

BLAST initiates extensions between sequences using a word, meaning that alignments need to share similarity along at least a 'word size' number of letters. The default value is 11 for nucleotide-nucleotide alignments and an exact match of 'word size' nucleotides between the two sequences is required; three is the default value for protein-protein matches and the sequences may merely be similar along the words, according to the matrix selected. If better sensitivity is needed, one should use a smaller value for the 'word size', but it is restricted to the range 7–20 for nucleotide comparisons and 2–3 for proteins.

The gapped BLAST 2.0 alignment algorithm uses dynamic programming to extend a central pair of aligned residues in both directions. This approach considers only the alignments that drop in score no more than 'x_dropoff' below the best score yet found. Increasing the 'x_dropoff' value increases the ability to produce a single gapped alignment rather than a collection of ungapped ones. Usually the value of 50 is enough to produce a single gapped alignment for an expected significance of 10, though these parameters may vary with the scoring system and gap costs. The expectation value is the expected frequency of the matches to be found merely by chance, according to the stochastic model of Karlin and Altschul [11]. To evaluate the statistical significance we choose to use the expectation value over the entire database search rather than the pairwise expectation. That makes it easier to compare the results of pairwise alignment with the results of the data base searches.

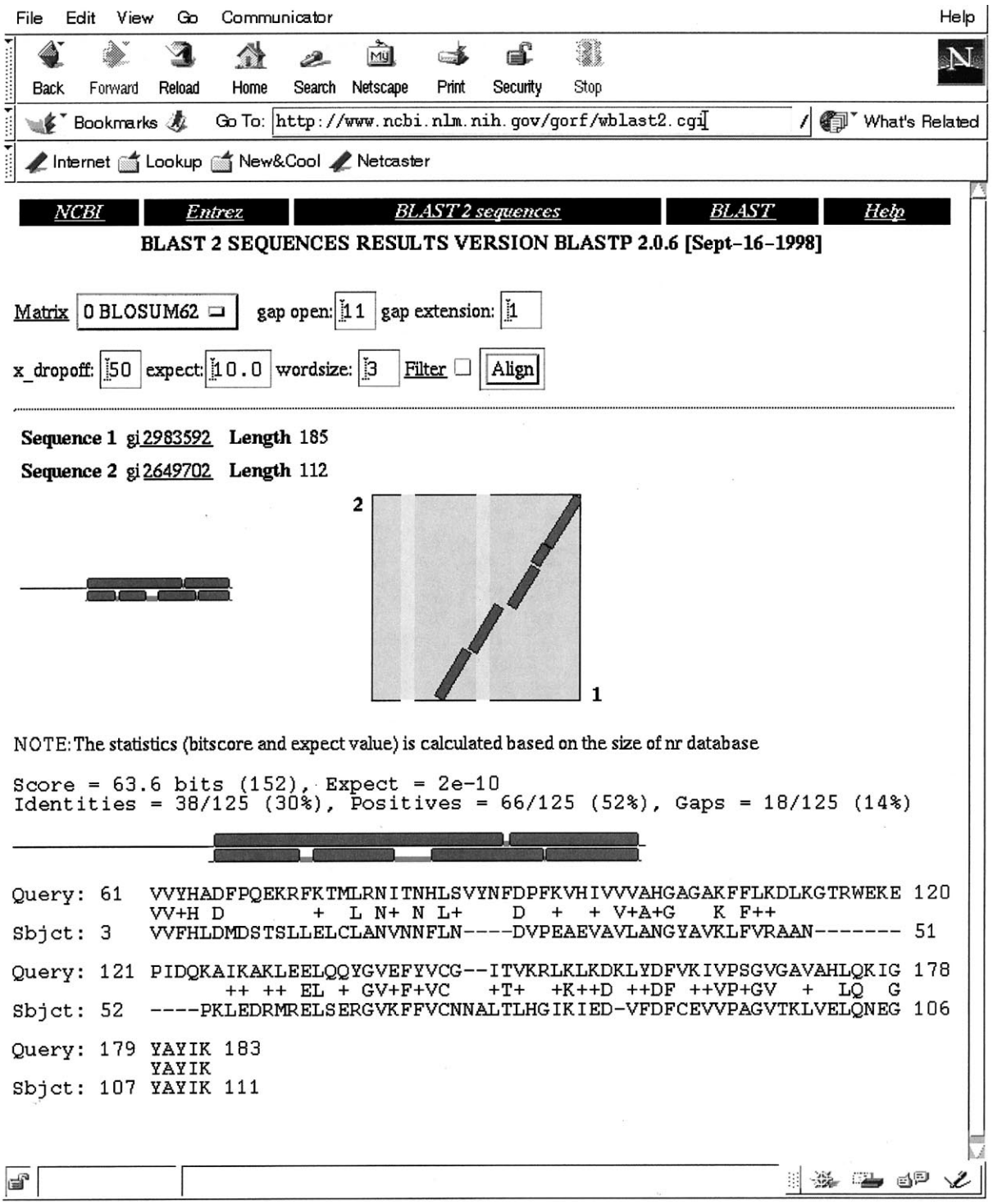

Fig. 1. The alignment of two protein sequences from *Aquifex aeolicus*. Blue rectangles: local alignment segments; red lines: gaps; light gray lines on dot plot picture represent the regions of low complexity on query sequence. The amino acids of query sequence shown in blue color will be filtered out if 'filter' option is applied.

2.3. Data constraints

The program is not generally useful for motif-style searching and aligning megabase size genomic sequences is not recommended. The maximum number of characters per sequence, that may be accommodated is ~150 kb, the optimal size of query sequence is about 1 kb.

3. Results and discussion

The result page (Fig. 1) starts with the values of parameters that were selected to produce the results. The user can recalculate the alignments by changing the parameters from this page and clicking on the 'Align' button, which provides a fast and convenient way of comparing the results for different values of parameters. It might be useful to compare the results of protein-protein alignments for different scoring matrices or change the expectation value.

The graphical representation shows schematically the set of gapped local alignments found between the two sequences with gaps shown in red. Clicking on the graphics brings the user to the detailed representation of that alignment described below. The dot plot picture provides the user with an overview of sequence similarity.

The detailed view for each segment of alignment includes the statistics with the percentage of identities, positives and gaps, schematic view, and the text alignment view. Note that the statistics are calculated based on the size of the NCBI non-redundant database. The bit score reported is the raw score converted to bits of information by multiplying by the Lambda parameter [11], with the raw score shown in brackets. The expectation value reports shows the number of alignments with that score one expects to find by chance. If the sequences are taken from the GenBank database and defined by gi or an accession number the hot link to the Entrez query system will be provided.

The last part of the report shows the parameters of the calculations and the summary of BLAST statistics. The report is the same as the regular BLAST report, providing an easy way to compare the results of the alignment of two sequences with the results of an entire database search.

Acknowledgments

We are grateful to David Lipman for numerous helpful discussions and to Roman Tatusov for valuable programming assistance.

References

[1] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. J. Mol. Biol. 215, 403–410.
[2] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.
[3] Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. J. Mol. Biol. 147, 195–197.
[4] Needleman, S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequences of two proteins. J. Mol. Biol. 48, 443–453.
[5] Tomb, J.-F., White, O., Kerlavage, A.R., Clayton, R.A., Sutton, G.G., Fleischmann, R.D., Ketchum, K.A., Klenk, H.P., Gill, S., Dougherty, B.A., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E.F., Peterson, S., Loftus, B., Richardson, D., Dodson, R., Khalak, H.G., Glodek, A., McKenney, K., Fitzegerald, L.M., Lee, N., Adams, M.D., Hickey, E.K., Berg, D.E., Gocayne, J.D., Utterback, T.R., Peterson, J.D., Kelley, J.M., Cotton, M.D., Weidman, J.M., Fujii, C., Bowman, C., Watthey, L., Wallin, E., Hayes, W.S., Borodovsky, M., Karp, P.D., Smith, H.O., Fraser, C.M. and Venter, J.C. (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature 388, 539–547.
[6] Alm, R.A., Ling, L.-S.L., Moir, D.T., King, B.L., Brown, E.D., Doig, P.C., Smith, D.R., Noonan, B., Guild, B.C., deJonge, B.L., Carmel, G., Tummino, P.J., Caruso, A., Uria-Nickelsen, M., Mills, D.M., Ives, C., Gibson, R., Merberg, D., Mills, S.D., Jiang, Q., Taylor, D.E., Vovis, G.F. and Trust, T.J. (1999) Genomic sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. Nature 397, 176–180.
[7] Henikoff, S. and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. USA, 89, 10915–10919.
[8] Altschul, S.F. and Erickson, B.W. (1986) Optimal sequence alignment using affine gap costs. Bull. Math. Biol. 48, 603–616.
[9] Karlin, S. and Altschul, S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. USA 87, 2264–2268.
[10] Wootton, J.C. and Federhen, S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. Comput. Chem. 17, 149–163.
[11] Altschul, S.F. and Gish, W. (1996) Local alignment statistics. Methods Enzymol. 266, 460–480.